# Re-humanizing the platform : Content moderators and the logic of care

Ruckenstein, Minna

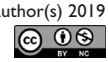
2020-06

Ruckenstein , M & Turunen , L 2020 , ' Re-humanizing the platform : Content moderators and the logic of care ' , New Media & Society , vol. 22 , no. 6 , 1461444819875990 , pp. þÿ 1 0 2 6 1 0 4 2 . h t t p s : / / d o i . o r g / 1 0 . 1 1 7 7 / 1 4 6 1 4 4 4 8 1 9 8 7 5 9 9 0

http://hdl.handle.net/10138/317845
https://doi.org/10.1177/1461444819875990

*Article*

new media & society
1–17



DOI: 10.1177/1461444819875990
journals.sagepub.com/home/nms


# Re-humanizing the platform: Content moderators and the logic of care

**Minna Ruckenstein**
**Linda Lisa Maria Turunen**
University of Helsinki, Finland

**Abstract**

With the goal of re-humanizing discussion platform operations, this study explores the knowledge and aims of commercial content moderators by reframing their work-related ideals through the notion of the "logic of care." In seeking to expand their professional realm by realigning users, moderators, and technical tools, moderators of discussion forums have turned to machines, ideally freeing up resources for real-time interaction between moderators and those who post. By focusing on care, the study calls for technical innovation that integrates moderators' aims with artificial intelligence systems. Rather than acknowledging human skills and resources in terms of moderation tools and discussion culture, the current platform logic forces moderators to operate like machines. Their discontent becomes understandable within a logic that diminishes their skills and vision. The moderator is left with assessing separate posts, rather than offering a meta-perspective to the discussion, overseeing and nurturing it.

**Keywords**

Artificial intelligence systems, content moderation, Finland, human–machine collaboration, ideal work, logic of care, platform

## Introduction

With growing concerns over misinformation and hate speech, discussion platforms face increasing political pressure to govern online communication. Simultaneously, however, conversations should be instantly public to strengthen the "real time" of social media

**Corresponding author:**
Minna Ruckenstein, Centre for Consumer Society Research and Helsinki Centre for Digital Humanities, P.O. Box 24, 00014 University of Helsinki, Finland.
Email: minna.ruckenstein@helsinki.fi

(Carmi, 2019; Kaplan and Haenlein, 2010). In the intersection of these diverse pursuits lies an area of work—content moderation—which is central for understanding the social and political tensions of current online cultures (Gillespie, 2018; Roberts, 2018, 2019). Recent research has examined content moderators as the hidden custodians of platforms, the unseen and silent guardians who maintain order and safety by overseeing visual and textual user-generated content (Gillespie, 2018; Klonick, 2017; Myers West, 2018; Roberts, 2016, 2019). We add to this approach a perspective that interrupts the connection between moderators' work and the dominant platform logic, exploring the knowledge and aims of commercial content moderators by reframing their work-related ideals with the notion of the "logic of care" elaborated by Annemarie Mol (2008).

The definition of platforms that guides this research is informed by Tarleton Gillespie's (2018) description of them as online services that "host, organize, and circulate users shared content or social interactions for them" (p. 18) and that "moderate the content and activity of users, using some logistics of detection, review and enforcement" (p. 21). Within these broad parameters, such platforms vary considerably, but what is crucial is the recurring logic with which they organize moderators' work. With the focus on care, we seek to go beyond that logic and re-humanize platform operations, making visible the human forces and ideals that the logic efficiently conceals (Carmi, 2019; Gray and Suri, 2019). In the context of content moderation, humans are involved in designing and implementing moderation software, also training machines and making decisions about online content. The move toward re-humanizing is a starting point for exploring the complexities of content moderation by re-establishing the human as a critical and creative actor in current and future platform arrangements.

We use the content moderators' perspective—their frustrations and discontents, and also their creativity and innovative thinking—to explore the insufficient and failed governance of platforms. By doing so, we expand the exploration of content moderation practices by arguing that the logic of care is central to moderators' understandings of how to govern online cultures productively. Mol (2008) outlines the logic of care as a different mode of thinking than the logic of choice, the dominant logic that we tend to follow when making judgments regarding autonomy and empowerment, and one that prevails when we choose from clearly demarcated options (Mol, 2008: 85). Significantly, the logic of care is not intrinsically better than the logic of choice. Yet, as an open-ended process, the logic of care is better equipped to deal with unpredictable events, flaws, and uncertainties (Mol, 2008: 95). This kind of open-endedness is particularly important in the current situation, characterized as it is by unresolved legal, political, and cultural tensions, with automated systems transforming the field of content moderation and introducing new kinds of human–machine associations (Gollatz et al., 2018).

We approach content moderation from its global margins with the aid of Finnish moderators—who mainly work for local companies—and suggest that their perspective tells a generalizable story of how current moderation practices distance moderators from the logic of care that they see as necessary for sustaining a healthy public culture. As we will argue, the logic of care materializes in moderation practices that fail and adjust, shape and reshape. In terms of content moderation, care allows moderation practices to go back and forth, to openly seek an outcome and a result. By building on our empirical material, we demonstrate that cultivating discussions is still at the heart of

how content moderation is imagined by professional moderators who care about their work and the future of online culture. From this perspective, commercial content moderation has similar aims as community moderation, seeking to support and nurture the online conversation with situated practices (Seering et al., 2019). This aspect of moderation is increasingly difficult to see, however, when content moderators are "trained to become algorithms, hidden from users and other actors" (Carmi, 2019: 454), their tasks are outsourced to workers in developing countries (Roberts, 2019), or the removal of undesirable content is delegated to artificial intelligence (AI) systems.

In the shadows of global platforms, Finnish moderators work for text-based online forums, owned by media companies, where Finns comment public affairs and talk about their daily struggles and interests. In the Finnish case, the moderation of textual material cannot be as easily outsourced to low-wage locations due to language specificity. Therefore, moderators are privileged in terms of decent incomes, and also closer to the life-worlds of writers: their work is often guided by an outspoken empathy for those who post. While the company puts economic pressure on moderators to constitute an efficient cleaning force, the moderators are keen to represent themselves as human interpreters by relating to and addressing the aims and needs of people online.

Max Weber (1949) famously drew the distinction between "real" and "ideal" types. The ideal type is formed of elements of a given phenomenon, but it is not meant to correspond to all of its real-life characteristics; rather, it underlines valued aspects of it. We argue that Finnish moderators have an ideal construct of work that helps them to navigate the social and economic reality of which they are part, and also see beyond it to what their work could be like, if it actually improved the conversational culture. By focusing on the perspective of content moderators rather than the platform or the user, we enter a scholarly realm where it matters how people think, feel, and act professionally and in organizations (Du Gay, 1997). As we suggest, advancing a perspective that calls for alternatives to current moderation practices demonstrates the fractured and failed character of current arrangements and draws attention to the ways in which they could be repaired. Here, we build on research that approaches failure as a new beginning, rather than as a breakdown or an endpoint. As Stephen Jackson (2014: 174), an STS scholar, asks, "what happens when we take erosion, breakdown, and decay, rather than novelty, growth, and progress, as our starting points in thinking through the nature, use, and effects of information technology and new media?" In our view, failure and associated brokenness provides a basis for repair work, ideally leading to more robust socio-technical innovation (Pink et al., 2018; Tanweer et al., 2016). Applying broken world thinking to content moderation pushes us to query the failures of moderation practices and consequent repair work, emphasizing the organic and incomplete nature of digital technologies. With the aid of content moderators, we are offered a perspective that bypasses the economic imperatives of targeted advertising and the policy fixes that promote accelerated content removal but fail to question the logic with which content moderation is practiced.

The article begins by describing our ethnographic data collection in detail, after which the theoretical frame guiding the paper—the logic of care—will be presented. Even if platforms circulate their policies concerning content moderation, they are reluctant to describe in detail the human and machine forces involved in practice (Roberts, 2018,

2019). Sharing overly detailed information about content moderation is considered a security risk, as moderation practices can be gamed, or circumvented (Gerrard, 2018). If knowledge about the lack of human moderators or their work conditions becomes general, it can harm the company brand. Thus, moderators tend to keep a low profile not only because of the non-disclosure agreements (NDAs) they have signed but also because they face threats both online and offline; consequently, we protect our interviewees by not revealing any identifiable information. Approaching content moderation work from various angles gave us a sufficiently comprehensive understanding of the field to allow us to look beyond individual moderators and identify challenges and struggles with current platform logics on a broader level. With the aid of the deep and diverse knowledge of the field, our informants' perspective highlights not only the insights gained in terms of content moderation work but also how, ideally, moderators could contribute to the development of online culture, both with and without machinic forces.

The division between the logic of care and the logic of choice inductively emerged from empirical data analysis when the practices of moderators were examined. We then introduced the framework to content moderators, to learn that it resonated with their aims. The closer the content moderators are, or have been, to online communities and everyday conversation flows, the more likely they are to criticize the effectiveness of their work. It is from this critical space that the moderators' understandings of ideal work emerge. When explicating this, they do not merely point out the limitations of current practices; rather, they would like to see them repaired. Thus, finally, we discuss how experienced moderators speculate on the ways human–machine collaborations could be used in imaginative and novel ways to produce better solutions for governing online conversations.

## Engaging with moderator generations

The empirical work that led to the framing of content moderation practices with the notion of logic of care started in Helsinki with the first author's participation in a master class for content moderators from fall 2015 to spring 2016. The class involved introductory lessons and actual moderation tasks, accompanied by free-flowing discussions with around 10 moderators. After laying the foundations for fieldwork among content moderators with the aid of the master class, we continued data gathering by means of participant observation and interviews. We have been able to experience how content moderation is done in practice, the processes and tools applied, and the kind of content addressed by commercial moderators. This sensitized us to the mental burdens of the work and the interpretative and evaluative dimension of moderation: to how difficult it is to accurately define "hateful speech" or the limits of "freedom of expression."

The second author conducted interviews in 2018 to deepen findings concerning content moderation, with purposively sampled informants (15), hand-picked with the aid of an experienced moderator. The number of experienced commercial content moderators is small in Finland and many of them have worked for the same platforms over a period of 16 years. By exploring content moderation within a longer time-frame, we could trace the development of moderation work. The informants' moderation-related careers varied from 2 to 15 years. The more senior among them might have started as online community

managers, but ended up working as content removers, or supervisors of other moderators due to the changing online environment and increasing levels of content. They were also part of the platforms' emerging trust and safety teams.

Most of our informants have experience with fairly large online platforms by Finnish standards (thousands to tens of thousands of messages daily). They have mainly been employed by platforms specializing in conversational fora, working with the moderation of anonymous textual content. One of the informants had only worked as a volunteer moderator, but because of the similarities between professional content moderation and community moderation, his long-term experience was useful for our study. The informants' experience covers the crafting of moderation policies, building online communities through community management, and developing technical tools for moderators. They have conducted content moderation practices manually, and some have been responsible for "training" AI. Interviewees who worked on the technical side of content moderation and platform development described automation processes in detail and critically assessed human–machine collaborations. Moderators who were merely applying and using the developed automation tools in everyday moderating practices viewed the machine in a more one-sided manner, mainly as a welcome aid in easing the monotonous removal task.

The opportunity to dive into the reflections and working experiences of moderators revealed the difference between moderator generations: moderators with long-term work experience started their careers at the beginning of millennium and had lived through the early days of social media; in comparison, those currently working in the field have a more distant relationship to their work and are more likely to perceive it as temporary. This difference led to the recognition that experienced moderators cared for the online conversations and platform developments in a way that the younger generation did not. Experienced moderators had played a part in building services and platforms in a nascent digital environment, working with its promises of participatory culture in ways that felt personally rewarding. They tend to have profound knowledge of platforms and an ability to see beyond current practices. With the understanding of how platforms are failing, they are eager to think of how they could be repaired. Our study gave them a venue to discuss development ideas that they had tried to promote within the companies where they worked, typically with insufficient response from the management.

In order to contextualize the perspective of content moderators further, we sought additional historical depth by interviewing three experts who pioneered the creation and development of online discussion platforms at the end of 1990s. For insights on future developments, we interviewed representatives from three Finnish companies offering AI services for content moderation. With the mounting public pressure to prevent and remove problematic content, companies are turning toward greater automation of content moderation and AI services routinely present their technologies as a catch-all solution to detect and filter hate speech and misinformation (Gollatz et al., 2018). We have closely followed Utopia Analytics, a Finnish company that specializes in automating content moderation, a service which has, for instance, been utilized in the largest Finnish online forum, Suomi24 (Finland24), since 2017. The company brands itself as having the aim to "bring democracy back into social media." The service that it offers is a text analytics AI technology which analyzes the corpus of data in

a given website, and then proceeds with the actual moderation tasks. At first, this takes place alongside human moderators who test and train the decisions of the system, and then the latter takes the role of supervisors, controlling machine-made decisions in controversial and less straightforward cases.

Our fieldwork culminated in a workshop where we presented findings on the moderation profession, the temporalities of moderation work, and human–machine associations to a group of 13 professional moderators. In addition to validating and clarifying our findings, the workshop turned out to be a welcome channel for peer support. Many experienced moderators were unemployed at the time, or seeking employment in other fields, underlining the fact that commitment to high-quality moderation is not a company priority. Content moderation tends to be treated as an extra cost rather than an opportunity to innovate and develop the platform. Machine forces are typically introduced as a replacement for human labor; automated tools are supposed to take care of removal at scale. From the perspective of labor studies, automation is regarded as replacing deskilled labor most effortlessly (Attewell, 1987), as it aids in the replicable aspects of work.

As we demonstrate, AI-enhanced moderation has become important for the way moderators imagine their work; "the machine" has gained a significant role in formulating ideal moderation work and the future logic of care. Yet, the machine is treated by moderators in a remarkably different way to that suggested by AI promoters, who sell the service of optimizing moderation tasks by delegating them to machines. Moderators are extremely critical of AI systems adopted with an economic logic to reduce human labor, emphasizing, like experts in the field, that the integration of such systems requires carefully designed and implemented human–machine collaboration (Gollatz et al., 2018). Time and again, moderators emphasize that AI systems cannot operate on their own, and that the final responsibility for moderation tasks should always be human-led. The simultaneous appreciation and critique of machine powers suggests that moderators approach technologies with caution and enthusiasm. They pursue collaboration between people and technical infrastructures but their work needs to be supported and enhanced with the tools that they are offered. With the focus on the logic of care, we are better equipped to recognize such collaborations, a theme to which we turn next. In light of AI-driven advancements, figuring out the appropriate division between machines and humans in terms of content moderation remains a pressing task. In terms of the logic of care, users and moderators, technologies and platforms, are all interrelated, and it is these relations that need attention and supervision.

## Content moderation and the logic of care

The logic of choice covers the selecting, deleting, and keeping of posted messages, while the logic of care is more far-reaching in its goals. The messages removed by moderators do not only have content that qualifies as illicit and offensive—the aspect receiving the most public attention; they may also be posted in the wrong place, contain private contact information, unsolicited advertising and marketing or spam, or be duplicates. The logic of care tackles all kinds of "mess" and "disorder" in the platforms, not merely

**Table 1.** Commercial content moderation in light of Mol's (2008) logic of care.

| Logic of choice | Logic of care |
|---|---|
| • Emphasizes the role of the decision-maker, either the human moderator or the machine.<br>• Covers the selecting, deleting, and keeping of posted messages.<br>• Moderation is solution-oriented: the unit of analysis is the message or the words in the message; there is a clear distinction between good and bad, morally right and wrong.<br>• Technologies are instruments; they are means to an end and the more effective these means are, the better.<br>• The logic of choice believes in policies and machine-led "one size fits all" services that can effectively conduct yes/no evaluation. | • Moderation is an interrelated and open-ended process, which requires interaction between the parties involved: users, moderators, and technologies are all participants in moderation.<br>• It covers community management, the selection and supervision of volunteer moderators, and stakeholder interaction: for instance, with advertisers, civil servants, or the police.<br>• Moderation is contextualized; the interpretation of messages is made in relation to the whole discussion, interaction, and history.<br>• It seeks to curate online discussions by educating writers about their own behavior and guiding them toward an improved dialogue. |

focusing on illegal or improper messages. Care is an ongoing process, aimed at maintaining and improving the overall platform culture.

While the logic of choice is fairly easy to pin down in terms of work practices, the logic of care consists of various kinds of relationalities that underline the entangled nature of moderation practices in terms of people and technologies (see Table 1). In light of the logic of care, moderation work is an ongoing process that covers tools and practices that go beyond individual moderators and their removal decisions. The logic of care remains platform-specific in the sense that it can involve work teams, volunteers, stakeholders, and technical tools in various ways, producing its own situated versions of care logics. For instance, moderation practices can involve flagging or deliberation by subcommittees of volunteers (Crawford and Gillespie, 2016).

The technical tools which Finnish moderators have employed over the years have typically been in-house: built to serve the needs of the moderators, with the resources at hand. Ideally, such tools make the context of the messages visible in terms of their content and who has posted them. While advancing the discussion culture, the logic of care requires responsive negotiation of what is suitable for a particular community (Seering et al., 2019); thus, content moderation is not a passive screening process, but continuous negotiation over desirable content. Here, the logic of care resonates with the notion that "communities evolve over time as a result of rule-breaking, rule-making and rule-enforcement" (Sternberg, 2012, cited by Seering et al., 2019: 4).

Based on our empirical material, we could detect a shift away from the logic of care in moderation practices. In the early days of social media, the moderators' aims were to attract users, spark online discussion, and guide it organically through participation: to build, care for, and curate the discussion culture. As discussion fora grew in popularity,

companies faced a new challenge; in addition to building desirability in order to generate new visits, they needed moderators to govern and control the produced material. Increasing numbers of messages generated an upsurge in inappropriate posts, and from the economic point of view, the desirability of a platform was dependent on keeping conversations clear of spam and offensive material.

With the flood of messages transforming the real-time participation of early content moderators in the direction of post-screening and more distant surveillance of online material, maintaining the moderators' role as caring professionals became more difficult to achieve. In one of the conversational platforms in Finland, signs of intensifying online harassment were visible around 2004–2005. Writers started posting, on a larger scale, content that was disturbing to moderators. In terms of upholding a good conversational space, the goal of the logic of care is to turn the negative into the more tolerable. It seeks to curate online discussions by educating writers about their own behavior and guiding them to an improved dialogue. Even if the ideal—healthier public discussion—might be difficult, if not impossible, to realize in the current online culture, it remains important for shared understandings of how content moderation should be practiced. As Mol (2008: 25) argues, "Care makes space for what is *not* possible."

## Committed to real-time conversation

Earlier research has recognized the importance of temporality in content moderation, discussing it in terms of pre-screening and post-screening (Gillespie, 2018; Roberts, 2016). The aim of pre-screening is to ensure that only ideal content is published, while editorial guidelines, policies, and rules governing the content posted shape and curate the discussion in advance. Inevitably, pre-screening blocks instantaneous discussion flow, as the messages are first reviewed, possibly also edited, and then published after the delay (Thurman, 2008). For a content moderator committed to fostering conversation spaces, pre-screening is regarded as blocking conversation as it inserts temporal hurdles in the ongoing discussion. As one of the moderators explains,

> With pre-screening, we would have no authentic discussion. The messages would arrive with a delay and the whole thread would be a mess: people would comment on the old stuff and then pre-screened messages would suddenly pop up.

As this comment makes clear, the emergent and organic nature of the conversation—its liveliness—should be actively protected. In light of the logic of care, "real time" is a characteristic of the online culture that moderators think needs to be a focus of attention, otherwise the discussion "regresses," as one of the moderators explains, and "becomes like an email list." If the discussions lose their temporal rhythm due to moderation decisions, the appeal of the online site, tying users to the continuous dynamic of the conversation, is lost. When explaining the characteristics of a well-curated platform, moderators emphasize monitoring the discussion flow rather than individual posts, suggesting that the posts need to be seen in the context of the discussion thread. Contextualization of the messages can lead to their contradictory valuation if provocative; from the perspective of the platform, provocation is valuable, as it generates discussion, more visits, and greater

user involvement. Return users link the moderation work to a platform's economic logic; indeed, users are the currency of platforms, an indication of success (Roberts, 2018). From the moderators' perspective, however, provocative messages are laborious as the discussion can quickly escalate and proceed in an inappropriate direction, requiring a moderator's presence.

The goal of maintaining the organic conversational rhythm underlines the rationale for moderators to separate their aims from those of traditional media, following the principles of pre-screening and editorial review. In addition, by distancing themselves from pre-screening, moderators distance their platforms from the regulation and cultural/political responsibility that restrict media publishers. Content moderators were vocal about this, highlighting their stance in the following manner: "If we did advance moderation, we, as a platform, would be responsible for all published content following the rules, policies and laws, as it would already have passed the review." The simultaneous commitment to the protection of conversational rhythms and the avoidance of pre-regulation suggests a professional positioning that cares for the online conversation at the same time as it avoids responsibility for content. Platforms have worked hard to maintain a public image of neutrality (Roberts, 2018), with moderators talking about freedom of expression to explain their position, thereby signaling the avoidance of excessive regulation. Yet, they also understand, through hands-on experience, the tensions and difficulties of limiting online content. As one of the moderators describes,

> It might not be easy to distinguish whether the message should be evaluated as a statement, an intention or a threat. Depending on how its motive is interpreted, the message is categorized as either illicit or appropriate—as it is legitimate to share opinions in a neutral way.

Extreme and hateful speech is constantly evolving and highly creative in that it can take into account company policies on what is understood to be proper. Word filters only work to a certain degree, because words can be written incorrectly to bluff the filter, for instance, with an extra space, or a hyphen. Furthermore, in some cases, the words per se might not be illicit, while their overall aim most certainly is. One of the interviewees describes a post with a detailed exploration of a painting picturing a prostitute. The automated content filtering removed the message, because it contained the word "whore" and the machine could not contextualize the message as the critical analysis of a painting. Instead of complaining about the removal, a new message from the same writer appeared shortly after with nearly the same content, but with the word "whore" replaced by "a naked woman." Obviously the writer knew why the message had been taken down, but wanted to make the point heard. With this example, the interviewee illustrates the role of users who continue to push against platforms, challenging and shaping them with their participation (Gerrard, 2018; Gillespie, 2018: 23).

Gillespie (2018: 197) argues that we have handed "the power to set and enforce the boundaries of appropriate public speech" to platform companies, but from the moderators' perspective, those boundaries are far from consistent (Carmi, 2019: 450). When discussing the difficulties of drawing the line with regard to appropriate content, the moderators point out that logic of choice is used for governing conversations: platforms behave like publishers when they manually remove content, or algorithmically delete,

downrank, and deprioritize it. At the same time, however, moderators admit the difficulties of applying a binary logic to public speech, especially when a machine is involved in the moderation process—think of irony, for instance (Nikunen, 2015). Nonetheless, platforms perform *as if* it were possible to maintain a consistent logic that selects what is posted, although this is a task that is close to impossible, which explains why platforms are so opaque and secretive when it comes to content moderation (Roberts, 2018).

## Efficient but distant moderators

As suggested above, inappropriate content could be material that is not posted according to the topical arrangement of the discussion forum—a post about cats when the subject is dogs, for example—whereupon such content is interpreted as harmful to platform order. In fact, talking about a wrong car type in a designated conversational area might generate a verbal war, underlining the importance of forum order to writers. Yet, during our fieldwork, moderators talked lengthily about the tensions associated with the removal of messages, noting that writers often find such action difficult to understand and accept:

> If a carefully crafted message is removed, the writer can swoop angrily in by email and request an explanation for the deletion. When a lot of time is spent on writing a message and the removal is not understood, it seems like eliminating it is unfair and requires an explanation.

Moderators justify why they do not comment on separate removal cases by referencing their workload. As Elinor Carmi (2019) describes, removal decisions are made "as fast as automated machines" (p. 450), within seconds. Working full-time as a content moderator for a discussion platform can involve reading, interpreting, contextualizing, and finally removing or keeping up to 1500 messages daily. Moderators underline that only messages that violate laws and company policies are deleted. As one of them puts it, "You need to be able to justify your removal decisions reliably, for anyone, anytime." Messages that require more evaluation are often reviewed with a team, in order to avoid decisions that cannot be defended. While retaining messages that are "neutral enough" means that a lot of forum content might be considered hateful or illicit by its readers, moderators might still underline that no false removal decisions should be made, depending on the platform. Others have internalized platform aims of speed and efficiency and argue that false removal decisions are inevitable. With the aid of AI, more possibilities are emerging for messages never to be published and undesirable content to be swiftly rejected. Services that seek removal at scale, proactively analyzing all uploaded content pre-publication, are bound to generate false positives, raising concerns about erroneous decision-making and over-policing of content (Gollatz et al., 2018).

A decade ago, a moderator could publicly explain why a message violated the rules of an online discussion. Often, after getting feedback, the writer either personally modified the message or deleted it altogether. The work of the moderator helped writers internalize the moderation principles and the logic was caring and educational, rather than punitive (see Myers West, 2018). In ideal conditions, content moderators made themselves unnecessary as cleaners; they no longer had to engage in removal work, as they had taught

writers to avoid emotional overheating and to handle their tensions in a civil manner. While caring for the conversational culture, the moderator did not have to focus on deleting messages but, rather, worked according to the principal that the online conversation was an ongoing process that required care and maintenance in order to flourish (Seering et al., 2019). The volunteer moderator explains, however, that this kind of care work is no longer possible due to the volume of messages: "In recent years, the group has become so huge that we need to keep it quick-and-dirty. Delete, without explaining in a detailed manner why it was removed. Everyone can read the rules, and hopefully follow them."

Over the years, technical tools have been developed to streamline screening and deleting problematic content: for instance, by automatically grouping, detecting or deleting messages with certain word combinations. Moderators rely on the efficiency of such tools, but express dislike for how automated screening targets only problematic content, with the result that it restricts the moderator's view of the conversational culture and limits the potential to curate it. "It is almost impossible to get a proper view of how the conversation is developing," one of the moderators explains. Another describes how the technical tool that their company uses exercises its own logic of choice and bundles together inappropriate messages, noting, "You only read those and not the 'real' conversation." Here, the experiences of moderators also vary, however, depending on the available tools. One of the moderators describes how the positive side of using the sorting tool is that you can structure the daily work according to message types: balancing the tasks of mass deleting and more detailed problem-solving.

Inevitably, technical tools distance moderators from conversations if they are not cleverly combined with the more encompassing aims of the logic of care. The moderators underline that the more distanced they are from the actual discussion, the less they have to offer as custodians of the platform. As one of them explains,

> When the moderation is conducted afterwards, only when the content is reported, you usually cannot influence the flow of discussion. Therefore, I see that the challenge is that post-screening fails to take into account the actual writers and contributors, who are the most important in terms of platform culture; instead, the work is done for readers or viewers. The contributors and writers are the critical group, which is why practices to support them should be developed.

Discussions are cleaned for readers and messages are deleted in order to maintain a purified public arena; however, the moderators lack mechanisms that would aid them in building a more self-aware conversational culture. Here, the moderators reference the transparency of moderation in much the same way as Gillespie (2018: 199), endorsing technical tools that would allow writers to follow how and when they post and why their messages are removed. Moderators believe that if their practices were made more visible—as maps or charts, for example—it would assist the efficient navigation of conversations that are "wild" or "clean," depending on the topic discussed. In conversational platforms, removed messages could, for instance, be used in generating "heat maps" to demonstrate which discussion areas, or topics, produce inappropriate content. If writers could see the patterned nature of discussion—how certain kinds of messages are repeatedly posted or deleted—perhaps it could have an effect on what people post; if not, at least it could generate important exchanges about the meta-characteristics of

an anonymous discussion. One of the main problems of current online culture is the impossibility of seeing the scale and spread of harmful content. A view, enabled by the data gathering of platform companies, of how coordinated harassment gathers momentum would be instructive for all readers and writers and allow a public intervention, or response in terms of condemning it.

Depending on the platform and topics discussed, the moderators point out that removed content is seldom more than 10% of all posted messages. Thus, the vast majority of human-generated posts meet the requirements of acceptable content. We were able to check this estimate by getting access to the percentages of removed messages on one of the biggest online forums in Finland. With the exception of a few specific discussion areas, in which the number of deletions was exceptionally high, removed messages were no more than 1–3% of all posted content. In terms of daily moderation practices, with a high volume of posted messages, these figures can still translate into great amounts of removed messages. Moderators value freedom of expression, but they realize that it creates problems and, once those are solved, others will arise. As with democracy, the manner of conducting online conversations continues to evolve through trial and error. Thus, the logic of care that moderators promote extends to provocative and heated messages that might be labor-intensive and require active participation and curation. It is those messages, however, that are central to adjusting the discussion culture.

## Hopes for human–machine collaboration

As suggested above, commonly used technical tools support the logic of choice, which aims to enhance the speed and efficiency of content removal, rather than the logic of care, with its goals of curating discussions. Gillespie (2018: 209) notes just how little technical innovation has supported user participation at the level of governance by inviting users either to participate in design processes or deliberate on values in terms of online culture. The same lack of innovation characterizes governance practices that content moderators could promote: better technical tools would offer a view to how discussions develop in terms of emotional or topical waves and how individual posts are distributed and by whom. Moderators need tools to combat those who purposefully game the system in order to harass others, repeatedly reposting the same messages.

As the moderator's work mainly focuses on the removal of negative and harmful content, it is distorting in terms of understanding what is going on within the platform. The technical arrangements naturalize the fact that moderators are confronted with brutal "dirt" and "waste" (Gillespie, 2018: 121), yet the mentally draining nature of moderation work is a direct consequence of this. Regularly exposed to cruel and vicious material, one of the moderators noticed that she was stricter with removal decisions at the beginning of her work shift than toward the end; emotionally, the work has a numbing effect and even though she is conscious of it, it is challenging to avoid it fully. In line with this, she emphasized that, at their best, AI-led solutions provide the consistency that a human reviewer lacks. As the promoters of automated moderation underscore, the machine tirelessly executes moderation tasks in a schematic and pedantic manner, following the rules derived from the training data with no exceptions.

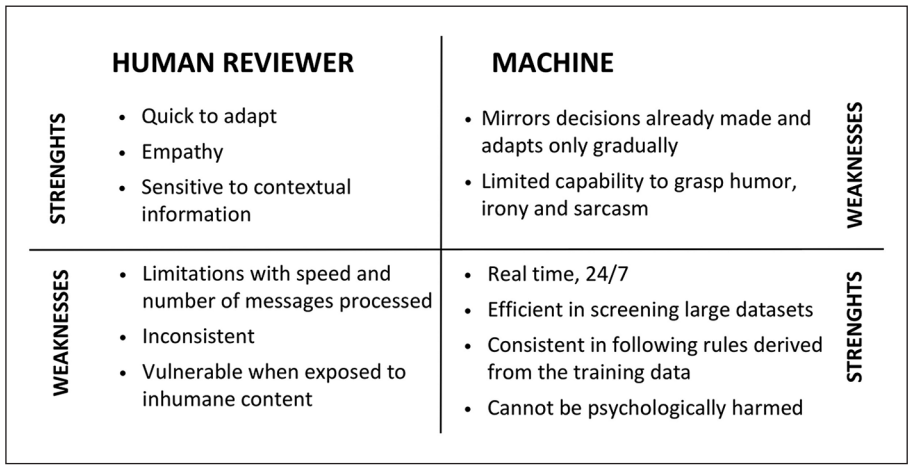



**Figure 1.** The strengths and weaknesses of human reviewers and machines compared.

Ultimately, however, AI is far from consistent. Even if a machine functions with precision, with a low margin of error, its quality as a content moderator can remain questionable. As one of our interviewees explains, the machine struggles with short and long messages. The short messages do not have sufficient relations between words for calculations, while negative content can disappear from long messages, because the machine counts how negative a message is overall; a single harmful or inappropriate element—quickly picked up by a human reviewer—is not enough to alert the machine. Because the machine focuses on relationships between words and calculates probabilities, if fails to recognize civilized insults and beautifully written dirt (Gröndahl et al., 2018).

When taken together, the strengths of machines that moderators recognize are related to the efficiency of the machine when exercising a logic of choice by following the rules derived from the training data (see Figure 1). Yet, the weaknesses of machines are similarly related to their dependence on training data. Moderators train the machine to perform in the future, enabling well-functioning real-time moderation practices to be realized. The machine gradually learns on the basis of moderators' past removal decisions and finally executes removal tasks according to existing training data. As one moderator describes, machine training requires the manual review of messages in order to produce a training dataset good enough for the machine to rely on:

> We are teaching the machine so that it can become more autonomous. Subsequently the human's role [in relation to the machine] becomes supervisory and the human can concentrate on participating in the discussion. The moderators' participatory role would then motivate users to discuss. That way, the moderator would not be in the background, but a visible actor in charge of the discussion.

The machine has its limitations with interpreting content; it has no empathy, sense of humor, or understanding of irony or sarcasm, merely mirroring decisions previously made by moderators. Machines slavishly execute removal tasks based on the training

data. When language and expressions change and new terms are adopted as inappropriate nicknames, moderators are quick to interpret, contextualize, and learn, but the machine only gradually adapts to the decisions made by human moderators, which means that the machine is not up to date. One of the moderators describes the "machine delay":

> In a way the machine is working with itself—with the examples of moderated messages that it has—but the moderators are working with users, so that the discussions stay clean. The machine reads the messages after the fact—it is made smart by human moderators—but it does not create anything new; it merely models the old. Human moderators indicate the direction to the machine; they are the important link and interpreter.

As this quote makes obvious, the moderator sees herself as a steward of the machine, in charge of the removal decisions rather than the machine, which can only be as capable as the examples that it is given. The machine cannot adapt to the online conversation, because they are not as steady and consistent as it would require. It can learn to mirror the probabilities of the conversation, but it does not care how the conversation develops, or what is at stake in terms of public culture.

## Conclusion

The goal of this study has been to re-humanize platform operations—a research move that is useful for exploring the complexities of online practices, including human–machine collaborations. We have argued that a focus on the logic of care in commercial content moderation interrupts the dominant platform logic by opening for exploration the aims and ideals that moderators have with regard to their work. With the goal of expanding their professional realm by realigning users, moderators, and technical tools, moderators have turned to machines as potential saviors of their work-related ideals. Unlike promoters of AI systems, portraying the machine as a replacement for human moderators, professional moderators maintain that automation transforms rather than substitutes human involvement in content moderation. In seeking improvements, moderators articulate an assisting role for the machine that evaluates the appropriateness of messages by following a pre-defined logic of choice. With the aid of AI-led systems, moderators could realign their work to apply the care logic and cultivate conversation by guiding the behavior of discussion participants, while the machine tirelessly sorts out dirt and waste, and deletes. Ideally, the machine frees up human resources for real-time interaction with writers online, which is labor intensive, but also more efficient for long-term aims. In other words, moderators express an ideal division of labor that delegates the tedious cleaning work to machines and frees moderators for more challenging work, consisting, for instance, of educating online communities and training and supervising machines.

When the moderators distinguish the labor of the human from that of the machine, they are, however, still attached to the current platform logic. The idea that machines exercise choice and humans engage in practices that align with the logic of care is limited in terms of rethinking content moderation. Instead, the quest for better technology should critique any pre-defined division of labor between the human and the machine. As Mol (2008) reminds us, the logic of choice and care cannot be entirely separated; both are

incorporated into, and mixed with, the practices of which they are elements. Instead of using the machine to root out illicit content—which always happens after a post has already appeared—machine-led or machine-enhanced solutions should be integrated into the work of supporting both moderators and users in the governance of the online culture. In order to disrupt the current logic, AI should become part of promoting practices of care in a more comprehensive manner, without ignoring the logic of choice. In line with this, the civic commitments of users, moderators, and platforms would have to become an integral part of technical initiatives. One way to work in this direction is to think about how moderators' work might be technically supported at the level of governance and with the orchestration of public culture.

With the goal of re-humanizing discussion platforms, we call for technical innovation that celebrates the strengths and merits of humans in relation to machines. Rather than acknowledging human skills and resources in terms of moderation tools and discussion culture, current platform logic forces the moderator to operate like a machine—"a human cleansing device" (Carmi, 2019: 451). The moderation tools slice the discussion into bits of text with the consequence that the moderator—rather than the machine—must slavishly sort through the unwanted and negative content. The current discontent of moderators and their quest for better technical tools becomes understandable within a platform logic that forces them to diminish their skills and vision as moderators and as humans. They are, thus, trapped in a cycle of responding to one post at a time rather than offering a meta-perspective to the discussion by overseeing and nurturing it. With the emphasis on re-humanizing the platform, our study calls for liberating content moderators from the machinic role to which they are assigned, and treating them as protagonists of past, present, and future online cultures.

## Acknowledgements

The authors warmly thank the moderators who contributed to the research. Special thanks to Pirjo Väyrynen. The two reviewers guided the rewriting of the article. Jesse Haapoja, Matti Nelimarkka, and Mari-Sanna Paukkeri (Utopia Analytics) offered suggestions and improvements when finalizing the piece.

## Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research has been funded by the Helsingin Sanomat Foundation and the Finnish Academy's Digital Humanities Programme's Citizen Mindscapes research consortium (2016–2019).

## ORCID iD


Minna Ruckenstein https://orcid.org/0000-0002-7600-1419

## References

Attewell P (1987) The deskilling controversy. *Work and Occupations* 14(3): 323–346.

Carmi E (2019) The hidden listeners: regulating the line from telephone operators to content moderators. *International Journal of Communication* 13: 440–458.

Crawford K and Gillespie T (2016) What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media and Society* 18(3): 410–428.

Du Gay P (1997) *Production of Culture/Cultures of Production*. London: SAGE.

Gerrard Y (2018) Beyond the hashtag: circumventing content moderation on social media. *New Media and Society* 20(12): 4492–4511.

Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven, CT: Yale University Press.

Gollatz K, Beer F and Katzenbach C (2018) The turn to artificial intelligence in governing communication online. Available at: https://www.ssoar.info/ssoar/handle/document/59528 (accessed 9 July 2019).

Gray ML and Suri S (2019) *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston, MA; New York: Eamon Dolan Books.

Gröndahl T, Pajola L, Juuti M, et al. (2018) All you need is "love": evading hate speech detection. In: *Proceedings of the 11th ACM workshop on artificial intelligence and security (AISec'18)*, Toronto, Canada, 15–19 October 2018, pp. 2–12. New York: ACM.

Jackson SJ (2014) Rethinking repair. In: Gillespie T, Boczkowski P and Foot K (eds) *Media Technologies: Essays on Communication, Materiality and Society*. Cambridge: MIT Press, pp. 221–240.

Kaplan AM and Haenlein M (2010) Users of the world, unite! The challenges and opportunities of social media. *Business Horizons* 53(1): 59–68.

Klonick K (2017) The new governors: the people rules, and processes governing online speech. *Harvard Law Review* 131: 1598–1670.

Mol A (2008) *The Logic of Care: Health and the Problem of Patient Choice*. London: Routledge.

Myers West S (2018) Censored, suspended, shadow banned: user interpretations of content moderation on social media platforms. *New Media and Society* 20(11): 4366–4383.

Nikunen K (2015) Politics of irony as the emerging sensibility of the anti-immigrant debate. In: Andreassen R and Vitus K (eds) *Affectivity and Race: Studies from Nordic Contexts*. New York: Routledge, pp. 21–42.

Pink S, Ruckenstein M, Willim R, et al. (2018) Broken data: conceptualising data in an emerging world. *Big Data & Society* 5(1):1–13.

Roberts ST (2016) Commercial content moderation: digital laborers' dirty work. In: Noble SU and Tynes B (eds) *The Intersectional Internet: Race, Sex, Class and Culture Online*. New York: Peter Lang, pp. 147–160.

Roberts ST (2018) Digital detritus: "Error" and the logic of opacity in social media content moderation. *First Monday* 23(3). Available at: https://firstmonday.org/ojs/index.php/fm/article/view/8283/6649

Roberts ST (2019) *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.

Seering J, Wang T, Yoon J, et al. (2019) Moderator engagement and community development in the age of algorithms. *New Media and Society* 21(7): 1417–1443.

Tanweer A, Fiore-Gartland B and Aragon C (2016) Impediment to insight to innovation: understanding data assemblages through the breakdown–repair process. *Information, Communication & Society* 19(6): 736–752.

Thurman N (2008) Forums for citizen journalists? Adoption of user generated content initiatives by online news media. *New Media and Society* 10(1): 139–157.

Weber M (1949) *Max Weber on the Methodology of the Social Sciences*. New York: The Free Press.

## Author biographies

Minna Ruckenstein works as an associate professor at the Centre for Consumer Society Research and the Helsinki Centre for Digital Humanities, University of Helsinki. She directs a research group that explores datafication by focusing on emotional, social, political and economic aspects of current and emerging data practices. The most recent collaborative project aims at re-humanizing automated decision making.

Linda Lisa Maria Turunen is a postdoctoral researcher at the Centre for Consumer Society Research, University of Helsinki. Her current research focuses on digital marketing and social media, outlining emerging human-machine work divisions.